Government Responses Matter:

Predicting Covid-19 cases in US using an empirical Bayesian time series framework

Ziyue Liu

Indiana University, Indianapolis, IN,USA, ziliu@iu.edu

AND

Wensheng Guo

University of Pennsylvania, Philadelphia, PA, USA, wguo@pennmedicine.upenn.edu

Abstract: Since the Covid-19 outbreak, researchers have been predicting how the epidemic will evolve, especially the number in each country, through using parametric extrapolations based on the history. In reality, the epidemic progressing in a particular country depends largely on its policy responses and interventions. Since the outbreaks in some countries are earlier than United States, the prediction of US cases can benefit from incorporating the similarity in their trajectories. We propose an empirical Bayesian time series framework to predict US cases using different countries as prior reference. The resultant forecast is based on observed US data and prior information from the reference country while accounting for different population sizes. When Italy is used as prior in the prediction, which the US data resemble the most, the cases in the US will exceed 300,000 by the beginning of April unless strong measures are adopted.

When facing an epidemic, people and government of a country may underestimate its seriousness in the beginning but will eventually step up their responses. Hence the case numbers tend to increase exponentially in the early stage, while the trends will gradually bend and plateau. Therefore, similarities in the case number trajectories can be observed in different countries, though the timing and severity can differ substantially due to different responses. Figure 1 displays the trajectories of total Covid-19 case numbers for China, S. Korea, Italy, France, Iran, Germany, Spain and USA using Johns Hopkins data. These countries have more days from time zero than US, where time zero is defined as first day with 100 or more (100+) cases as a heuristic but widely used choice[1]. The curve of South Korea increased rapidly early on but quickly bended and plateaued, for which S. Korea's swift and deterministic policy responses are credited[2]. China exhibits similar but later flattening pattern, which agrees with its missing early intervention window, but later extreme lockdown policy implementation[3]. On the other hand, the cases in Italy and France have grown exponentially until recent days, which have partially been attributed to their late and weak policy responses[4]. The US trajectory is almost linear on the logarithm scale. While the US government is catching up with policies such as work/study from home, social distancing and self-quarantine, the effect has not seen in the trajectory.

Existing Covid-19 forecasting are extrapolations into the future time[5-11]. Their validity relies on the crucial but unrealistic assumption that the future trajectories are completely determined by the history. This by design cannot incorporate government responses yet to come. Not surprisingly, these predictions can be off the target. For example, Fanelli and Piazza[7] predicted a maximum number of cases in Italy to be 15,000, where the real cases have already multipled. Batista[8] predicted the pandemic should peak around Feb 9th, 2020, but it shows no sign of slowing down into late-March, 2020. Zheng et al[9] predicted about 20,000 cases in South Korea, which is unlikely to happen given its current flat trend around 9,000. Models used in

these forecasting are mainly the susceptible-infected-removed (SIR) models and its variants[5-8]. Others include state transition model[9], parametric growth curve models such as logistic curves[10], and auto regressive integrated moving average (ARIMA) models[11].

Proposed Methods

We propose an empirical Bayesian time series framework to forecast the US trajectory by utilizing the idea of internal time. Since the virus spread to different countries at different time, their trajectories are different in calendar time but comparable in internal time. We define time zero as the first day with 100 or more cases in a given country. We first model the trajectories of the eight countries by a functional mixed effects model[12], where different countries shared a similar mean trajectory over time, and each country has its own random deviation curve. An additional scalar fixed effect parameter is incorporated to account for different population sizes on the natural logarithm scale. The estimated coefficient is 0.34, suggesting while the population size has some effects on the cases numbers, it is not fully proportion to the population. Both the population-average curve and random deviation curves are modeled by cubic splines. The model is then casted into state space model for computational efficiency and forecasting. The smoothing parameters and the variances are estimated through maximum likelihood.

Based on the estimated parameters using the eight countries, our next task is forecast the US cases while incorporate one of the countries as the prior information. This is done through constructing conditional state space model from the functional mixed effects model conditional on the observed data of the specified country[13]. By running the Kalman filter forward on the conditional state space model with the US time series data and into the future, the results are the posterior prediction incorporating both the prior information from the specific country and the observed US data. As the reference country is only specified as the prior, the posterior can be substantially different from the prior, suggesting strong deviation from the reference country.

In addition, the observed US data can be substantially different from posterior prediction, indicating that the US case are following a different trajectory because of different policy responses. More technical details are given in the Supplement.

Data Analysis

The Johns Hopkins University CSSE data were downloaded from its GitHub repository (https://github.com/CSSEGISandData/COVID-19). We modeled the natural logarithms of the case numbers as the outcome. The data were then used for prediction using the proposed method. After the posterior means and variances were calculated and the 95% prediction intervals were constructed, they were taken exponential to transform back to the original scale. The whole data analysis from reading in the data to plotting the results took less than 10 seconds on personal computer with Intel® Core™i76600U CPU @ 2.60GHz, 2801Mhz, 2 Cores, 4 Logical Processors.

Results

Results based on US data up to March 26th, 2020 are shown in Figure 2~4. Two important observations can be made from these figures. There is no apparent slowing down yet for US trajectory based on either the observed trend or predicted trend. This indicates that US is still in its exponentially increasing phase in the near future.

Figure 2 displays the results using Italy as prior. It shows that US and Italy have similar patterns and majority of the observed US data are in the 95% prediction intervals. This suggests that the trajectory in Italy serves as a good prior for the US prediction. Based on this prediction, on the next day as March 27th, 2020, US may have as many as 108,595 cases. In about 10 days, the US case number will exceed 300,000 around April 4th, 2020 shall the US policy responses have similarly effects as Italy.

Figure 3 displays the results using China as prior. It shows that the observed US case numbers are already higher than the predicted values. Even if the US policy responses have similar effect as China, US case numbers will exceed 150,000 around April 11th, 2020. The results using South Korea as prior are displayed in Figure 4. US case numbers are predicted to exceed 200,000 around April 6th, 2020. Since the observed US data are already well above the upper bound of the 95% prediction intervals, the data from China and South Korea are not good priors for the US prediction, suggesting that the situation in the US will be much worse than those in China and South Korea.

Conclusion

We have proposed a new prediction method for predicting total COVID-19 cases of US by incorporating the information from other countries. While we demonstrated our method in predicting US cases, our method can be used for predicting state-by-state data as well as hospital-by-hospital data. Our prediction intervals are much smaller than most exiting methods due to the additional information from the reference country. We show that the current trajectory in US is most similar to that in Italy. The stronger response from Italy has led to slowing down of the spread in the last few days, while the effect of social distancing in the US has not shown in the observed data.

It is well-known that there are serious under-reporting or under-detection of cases in various countries and under-reporting rates may be very different across counties. This can contribute to substantial differences in the trajectories. With the advance of testing techniques, more and more people are tested in the US. This may also explain why the reported cases in the US are substantially higher than other countries in the same stages.

References

1. Rattner, N. & Schoen, J. W. These charts show how fast coronavirus cases are spreading – and what it takes to flatten the curve. *CNBC* (2020). Available at: https://www.cnbc.com/2020/03/22/these-charts-show-how-fast-coronavirus-cases-are-spreading.html
2. Zastrow, M. South Korea is reporting intimate details of COVID-19 cases: has it helped? *Nature News* (2020). Available at: https://www.nature.com/articles/d41586-020-00740-y
3. Cyanoski, D. What China's coronavirus response can teach the rest of the world? *Nature News* (2020). Available at: https://www.nature.com/articles/d41586-020-00741-x
4. Donadio, R. Italy's coronavirus response is a warning from the future. The Atlantic, March 8th, 2020. https://www.theatlantic.com/international/archive/2020/03/italy-coronavirus-covid19-west-europe-future/607660/
5. Zhang, Y. et al Prediction of the COVID-19 outbreak based on a realistic stochastic model. Preprint at https://doi.org/10.1101/2020.03.10.20033803
6. Wang, L. et al. An epidemiological forecast model and software assessing interventions on COVID-19 epidemic in China. Preprint at https://doi.org/10.1101/2020.02.29.20029421
7. Fanelli, D. & Piazza, F. Analysis and forecast of COVID-19 spreading in China, Italy and France. Preprint at https://arxiv.org/abs/2003.06031
8. Batista, M. Estimation of the final size of the COVID-19 epidemic. Preprint at https://doi.org/10.1101/2020.02.16.20023606
9. Zheng, Z., Wu, K., Yao, Z., Zheng, J. & Chen, J. The prediction for development of COVID-19 in global major epidemic areas through empirical trends in China by utilizing state transition matrix model. Preprint at https://doi.org/10.1101/2020.03.10.20033670
10. Buizza, R. Probabilistic prediction of COVID-19 infections for China and Italy, using an ensemble of stochastically-perturbed logistic curves. Preprint at https://arxiv.org/abs/2003.06418

11. Benvenuto, D., Giovanetti, M., Vassalo, L. & Angeletti, Silvia. Application of the ARIMA model on the COVID-2019 epidemic dataset. Preprint at https://doi.org/10.1016/j.dib.2020.105340

12. Guo, W. Functional mixed effects models. *Biometrics*, **58**, 121-8 (2002).

13. Guo, W. Dynamic state space models. *Journal of Time Series analysis*, **24**, 149-158 (2003).

14. Durbin, J. & Koopman, S.J. *Time Series Analysis by State Space Methods* (2nd edn). Oxford University Press: Oxford, UK. (2012).

Figure 1. Cases numbers for China, S. Korea, Italy, France, Spain, Germany, Iran, and US on the natural logarithm scale. For the first seven countries, the raw data are shown as symbols, the smoothed trends as solid lines, and the 95% confidence intervals in dotted lines. For US, only the raw data are displayed.

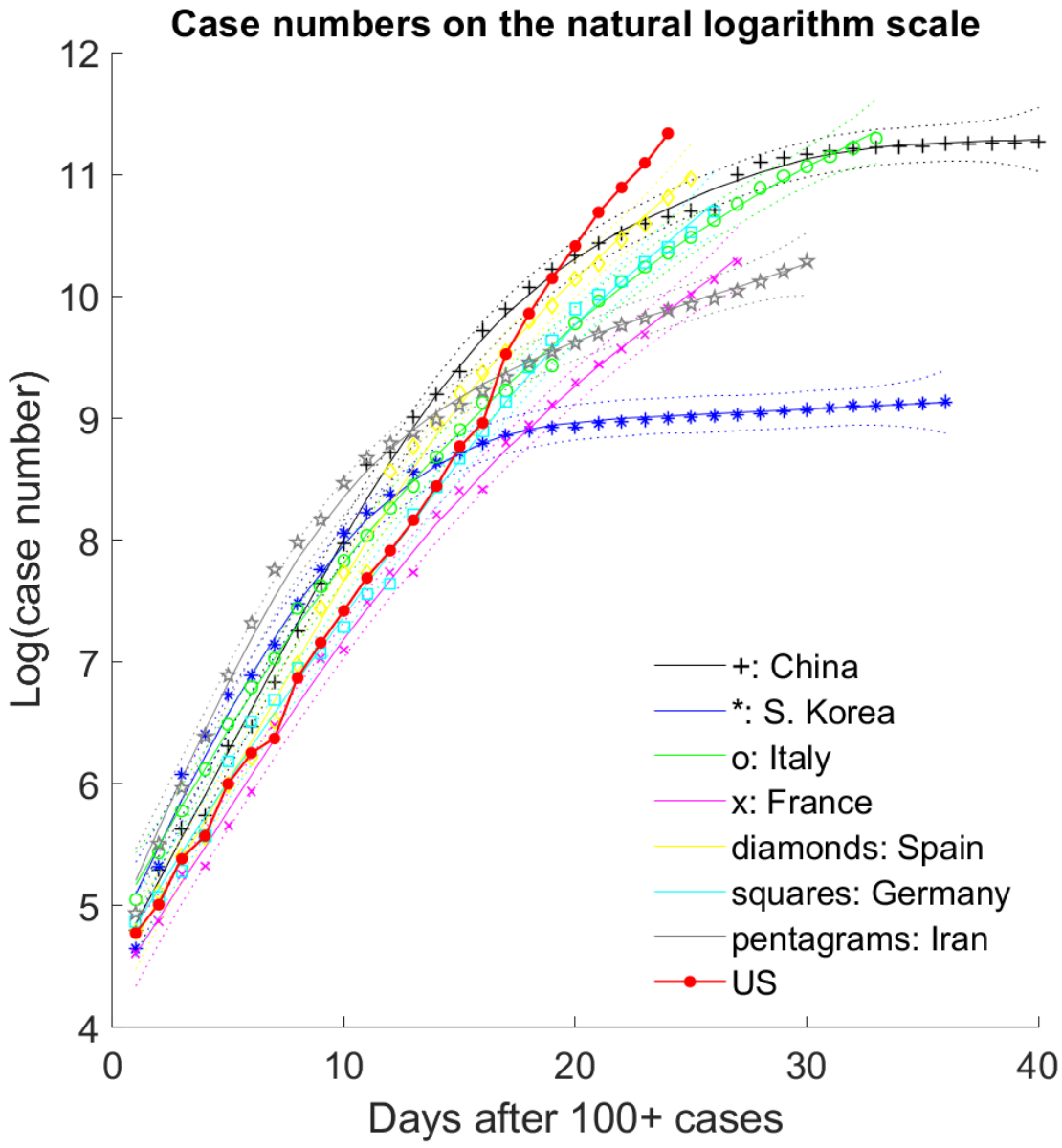


Figure 2. Prediction of the US trajectory using Italy as prior. The horizontal axis is based on US calendar time. On April 4th, 2020, the mean prediction value exceeds 300,000.

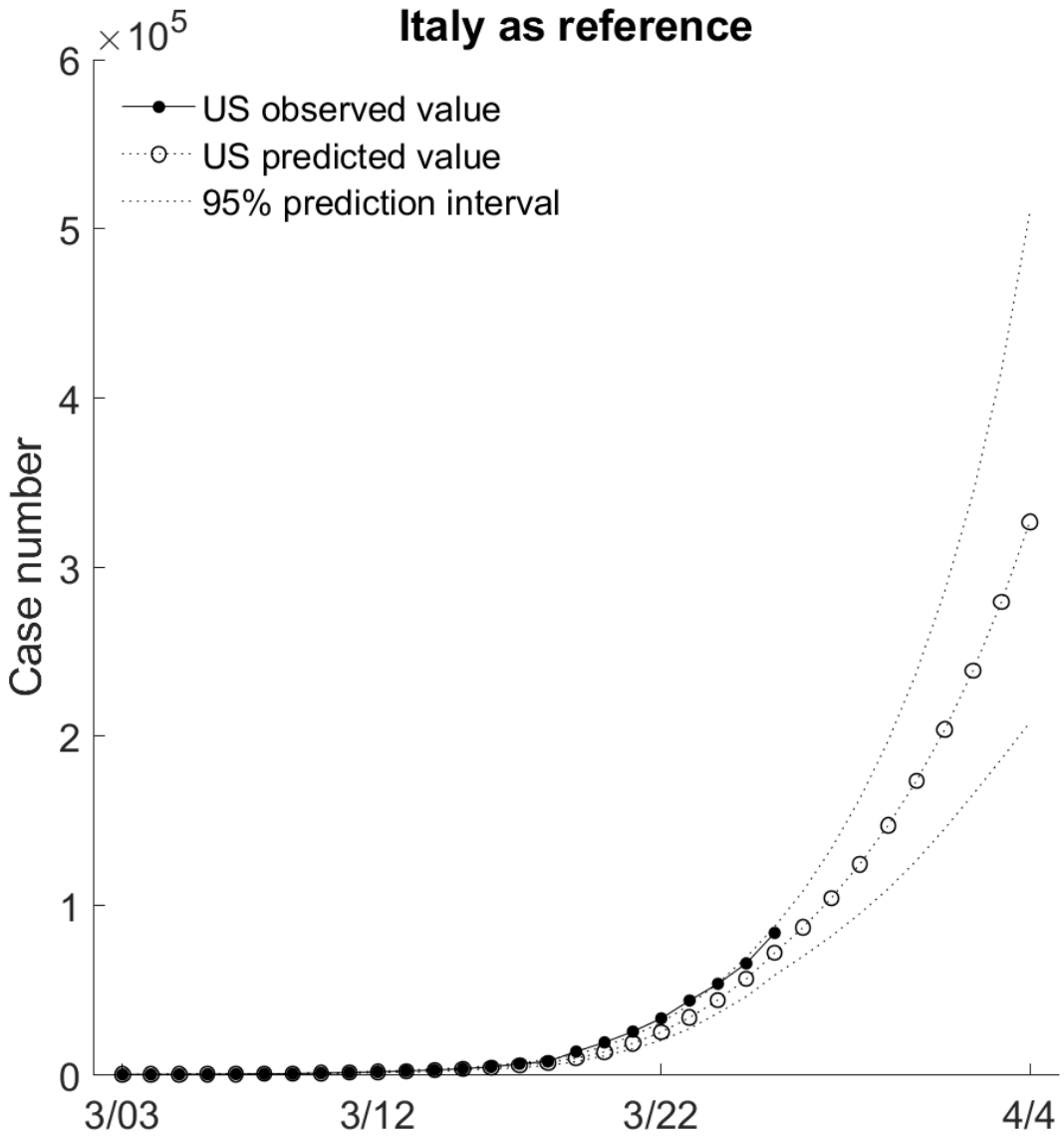


Figure 3. Prediction of the US trajectory using China as prior. The horizontal axis is based on US calendar time. On April 11th, 2020, the mean prediction value exceeds 150,000.

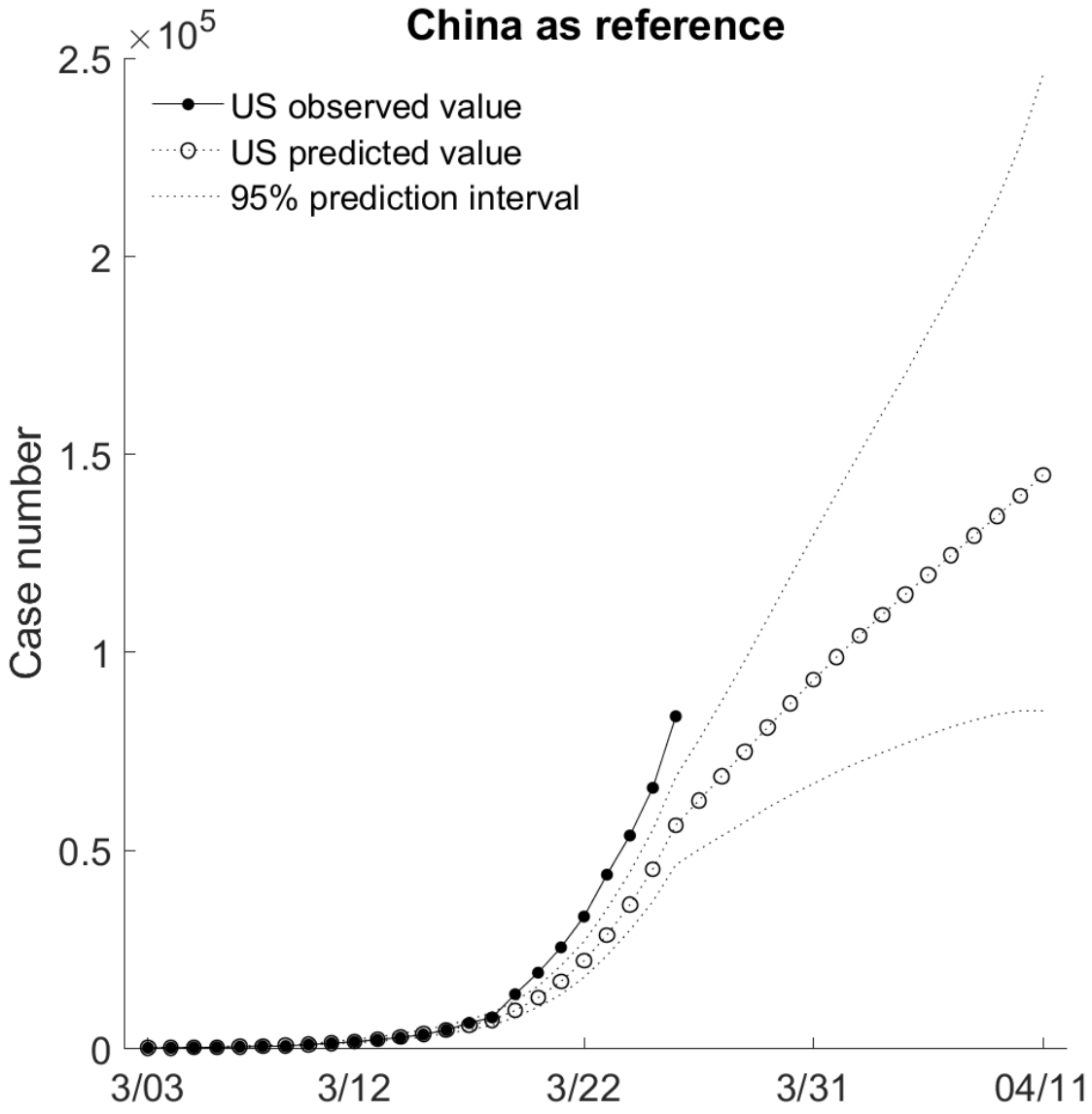
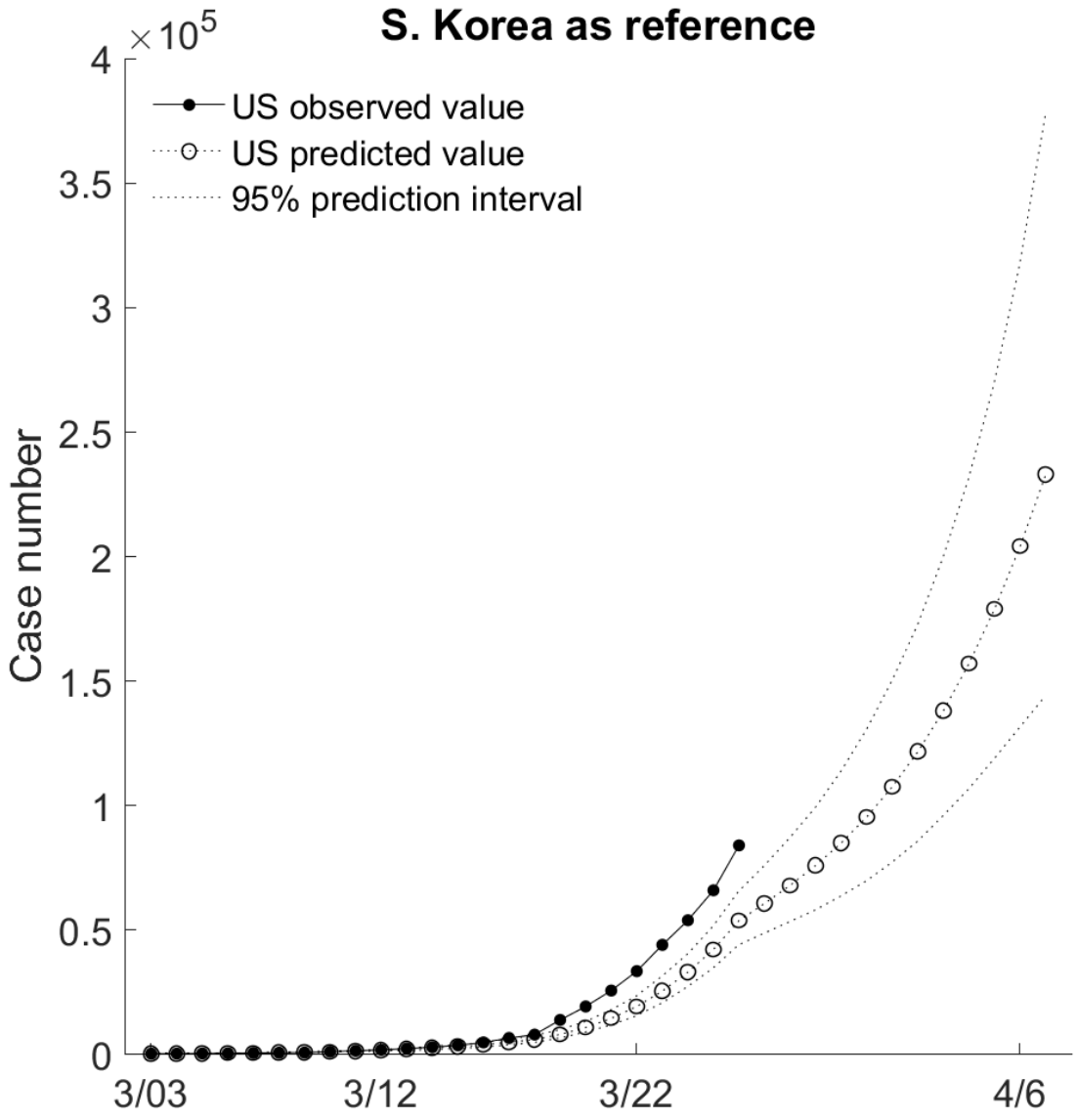


Figure 4. Prediction of the US trajectory using S. Korea as prior. The horizontal axis is based on US calendar time. On April 6th, the mean prediction value exceeds 200,000.



Supplement

Functional mixed effects model

Let N_{ij} be the number of total COVID-19 cases for the i^{th} country on the j^{th} day, where day 1 is defined as the first day with 100 or more cases. We model the natural logarithm of N_{ij}, $y_{ij} = \log(N_{ij})$, by a functional mixed effects model[1] as

$$y_{ij} = \beta \log(p_i) + f(t_j) + g_i(t_j) + \varepsilon_{ij}, \quad (1)$$

Where p_i is the population size in the unit of millions, β is the fixed effect slope for $\log(p_i)$, $f(t_j)$ is the functional fixed effects, $g_i(t_j)$ is the functional random effects, and $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma_e^2)$ is the error term. We model $f(t_j)$ by a cubic smoothing spline with the state space representation as

$$\begin{pmatrix} f(t_j) \\ f'(t_j) \end{pmatrix} = H_f \begin{pmatrix} f(t_{j-1}) \\ f'(t_{j-1}) \end{pmatrix} + \boldsymbol{\eta}_j, \quad (2)$$

where $f'(t_j)$ is the first derivative with respect to time. The state transition matrix $H_f = \begin{pmatrix} 1 & \delta t \\ 0 & 1 \end{pmatrix}$ with δt is the time interval between two points with the overall time range scaled to $[0,1]$. The state innovation vector $\boldsymbol{\eta}_j \sim \mathrm{N}(\mathbf{0}, \Sigma_f)$, with $\Sigma_f = \lambda_f^{-1} \begin{pmatrix} \delta t^3/3 & \delta t^2/2 \\ \delta t^2/2 & \delta t \end{pmatrix}$ and λ_f is the smoothing parameter. The state vector $\begin{pmatrix} f(t_j) \\ f'(t_j) \end{pmatrix}$ is initialized at time zero as $\begin{pmatrix} f(0) \\ f'(0) \end{pmatrix} \sim \mathrm{N}(\mathbf{0}, \kappa I)$ with $\kappa \to \infty$ and I is the identity matrix. We model $g_i(t_j)$ similarly but using a sine function in the W_0 space with the state space representation[2] as

$$\begin{pmatrix} g_i(t_j) \\ g'_i(t_j) \end{pmatrix} = H_g \begin{pmatrix} g_i(t_{j-1}) \\ g'_i(t_{j-1}) \end{pmatrix} + \boldsymbol{\xi}_{ij} \quad (3)$$

with time rescaled to $[0, 0.5]$. The state transition matrix $H_g = \begin{pmatrix} \cos(2\pi\delta t) & \frac{1}{2\pi}\sin(2\pi\delta t) \\ -2\pi \sin(2\pi\delta t) & \cos(2\pi\delta t) \end{pmatrix}$.

The state innovation vector $\boldsymbol{\xi}_{ij} \sim \boldsymbol{N}(\mathbf{0}, \Sigma_g)$, with

$$\Sigma_g = \lambda_g^{-1}\begin{pmatrix} \frac{1}{8\pi^2}\delta t - \frac{1}{32\pi^2}\sin(4\pi\delta t) & \frac{1}{16\pi^2}(1-\cos(4\pi\delta t)) \\ \frac{1}{16\pi^2}(1-\cos(4\pi\delta t)) & \frac{1}{8\pi}\sin(4\pi\delta t) + \frac{\delta t}{2} \end{pmatrix}, \quad 4$$

and λ_g is the smoothing parameter. The state vector $\begin{pmatrix} g_i(t_j) \\ g'_i(t_j) \end{pmatrix}$ is initialized at time zero as

$\begin{pmatrix} g_i(0) \\ g'_i(0) \end{pmatrix} \sim \mathrm{N}\left(\mathbf{0}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}\right)$.

Parameter estimation.

Data from eight countries (China, S. Korea, Italy, France, Iran, Germany, Spain and USA) were used. Let n_i denote the number of observations for subject $i = 1, \ldots, k$, $\boldsymbol{y}_i$ the corresponding observed data vector, $\boldsymbol{t}_i$ the time vector, $\boldsymbol{f}(\boldsymbol{t}_i)$ the vector of fixed functional effects evaluated at $\boldsymbol{t}_i$, $\boldsymbol{g}_i(\boldsymbol{t}_i)$ the functional random effect, $f"(t)$ and $g"_i(t)$ the second derivatives with respect to time, and $\boldsymbol{y}$ the overall observed data vector. The following penalized log-likelihood[3] was maximized to estimate the parameter vector $\boldsymbol{\theta} = (\beta, \lambda_f, \lambda_g, \sigma_1^2, \sigma_2^2, \sigma_e^2)$

$$l(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{k} -\frac{1}{2\sigma_e^2}\left(\boldsymbol{y}_i - \beta\log(p_i) - \boldsymbol{f}(\boldsymbol{t}_i) - \boldsymbol{g}_i(\boldsymbol{t}_i)\right)^2 - \lambda_f \int f"(u)du \quad 5$$

$$- \lambda_g \sum_{i=1}^{k} \int g"(u)du.$$

The maximum likelihood parameter estimates were $\widehat{\boldsymbol{\theta}} = (0.34, 306.69, 0.09, 1.40, 5.81, 0.05)$**.** We adopt an empirical Bayes approach such that these parameters are treated as known in the following steps.

Construction of the conditional SSM.

For the i^{th} reference country, the conditional SSM was constructed on the state vectors dimension of six as $\boldsymbol{x}_{t_j} = \left(f(t_j) \quad f'(t_j) \quad g_i(t_j) \quad g'_i(t_j) \quad g_{US}(t_j) \quad g'_{US}(t_j)\right)^T$, where subscript 'US' denote US-specific component. The working data are $\tilde{y}_{ij} = y_{ij} - \hat{\beta}\log(p_i)$. The observation

matrix is $F = (1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0)$, the state transition matrix is $H = \text{block diagonal}\left(H_f, H_g, H_g\right)$, the state innovation vector $\boldsymbol{\varpi}_j = \left(\left(\boldsymbol{\eta}_j\right)^T \quad \left(\boldsymbol{\xi}_{ij}\right)^T \quad \left(\boldsymbol{\xi}_{US,j}\right)^T\right)^T$ distributed as $\boldsymbol{\varpi}_j \sim \mathrm{N}\left(\mathbf{0}, \Sigma_j\right)$ where $\Sigma_j = \text{block diagonal}\left(\Sigma_f, \Sigma_g, \Sigma_g\right)$. The working SSM is

$$\begin{aligned} \tilde{y}_{ij} &= F\boldsymbol{x}_{t_j} + \varepsilon_{ij}, \\ \boldsymbol{x}_{t_j} &= H\boldsymbol{x}_{t_{j-1}} + \boldsymbol{\varpi}_j. \end{aligned} \tag{6}$$

Let $\boldsymbol{m}(\cdot)$ denote the mean and $W(\cdot)$ the variance of $\boldsymbol{x}_{t_j}$. The conditional SSM was constructed in three steps using the dynamic state space models[4].

Step 1. The forward filtering: for $j = 1, \ldots, n_i$

$$\begin{aligned} \boldsymbol{m}(j|j-1) &= H\boldsymbol{m}(j-1|j-1), \\ W(j|j-1) &= HW(j-1|j-1)H^T + \Sigma_j, \\ e_j &= \tilde{y}_{ij} - F\boldsymbol{m}(j|j-1), \\ K_j &= W(j|j-1)F^T, \\ \boldsymbol{m}(j|j) &= \boldsymbol{m}(j|j-1) + \frac{K_j e_j}{(FW(j|j-1)F^T + \sigma_e^2)}, \\ W(j|j) &= W(j|j-1) - \frac{K_j K_j^T}{(FW(j|j-1)F^T + \sigma_e^2)}. \end{aligned} \tag{7}$$

Step 2. The backward smoothing: at $j = n_i$ let $\boldsymbol{m}(j|n_i) = \boldsymbol{m}(n_i|n_i)$ and $W(j|n_i) = W(n_i|n_i)$, for $j = n_i - 1, \ldots, 0$

$$\begin{aligned} C_j &= W(j|j)H^T\{W(j+1|j)\}^{-1}, \\ \boldsymbol{m}(j|n_i) &= \boldsymbol{m}(j|j) + C_j\{\boldsymbol{m}(j+1|n_i) - \boldsymbol{m}(j+1|j)\}, \\ W(j|n_i) &= W(j|j) + C_j\{W(j+1|n_i) - W(j+1|j)\}C_j^T. \end{aligned} \tag{8}$$

Step 3. Construction step: let $\tilde{\boldsymbol{x}}_0 \sim \mathrm{N}\left(\boldsymbol{m}(0|n_i), W(0|n_i)\right)$, for $j = 1, \ldots, n_i$

$$\begin{aligned} \tilde{H}_j &= W(j|n_i)C_j^T\{W(j-1|n_i)\}^{-1}, \\ \tilde{\boldsymbol{\mu}}_j &= \boldsymbol{m}(j|n_i) - \tilde{H}_j\boldsymbol{m}(j-1|n_i), \\ \tilde{\Sigma}_j &= W(j|n_i) - \tilde{H}_j W(j-1|n_i)\tilde{H}_j^T. \end{aligned} \tag{9}$$

where $\tilde{H}_j$ is the state transition matrix, $\tilde{\boldsymbol{\mu}}_j$ is the mean vector and $\tilde{\Sigma}_j$ is the variance matrix of the state innovations.

Predict US trajectory.

The US trajectory was predicted by running the n_{US} observations $\tilde{y}_{US,j} = y_{US,j} - \hat{\beta}\log(p_{US})$ through the conditional SSM constructed above with an observation matrix $F = (1 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0)$. For $j = 1, \dots, n_{US}$, the first two equations in (7) generated the one-step ahead prediction, while the last two equations generate the filtered values. Further running the model for $j = n_{US} + 1, \dots, n_i$ generated predictions into the future time for US. $\hat{\beta}\log(p_{US})$ was then added back to recover the effects of population size.

References for the supplement

1. Guo, W. Functional mixed effects models. *Biometrics*, **58**, 121-8. (2002).
2. Qin, L. Functional models using smoothing splines, a state space approach. *Dissertation*.
3. Wahba, G. A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics*, **13**, 1378-1402. (1985).
4. Guo, W. Dynamic state space models. *Journal of Time Series analysis*, **24**, 149-158. (2003).